Synonymous codon usage in Bacillus subtilis reflects both translational selection and mutational biases

Volume 15 Number 19 1987 Nucleic Acids Research

Synonymous codon usage in *Bacillus subtilis* reflects both translational selection and mutational biases

Denis C.Shields and Paul M.Sharp*

Department of Genetics, Trinity College, Dublin 2, Ireland

Received July 16, 1987; Revised and Accepted September 4, 1987

ABSTRACT

Codon usage data for 56 Bacillus subtilis genes show that synonymous codon usage in B.subtilis is less biased than in Escherichia coli, or in Saccharomyces cerevisiae. Nevertheless, certain genes with a high codon bias can be identified by correspondence analysis, and also by various indices of codon bias. These genes are very highly expressed, and a general trend (a decrease) in codon bias across genes seems to correspond to decreasing expression level. This, then, may be a general phenomenon in unicellular organisms. The unusually small effect of translational selection on the pattern of codon usage in lowly expressed genes in B.subtilis yields similar dinucleotide frequencies among different codon positions, and on complementary strands. These patterns could arise through selection on DNA structure, but more probably are largely determined by mutation. This prevalence of mutational bias could lead to difficulties in assessing whether open reading frames encode proteins.

INTRODUCTION

In many genes from many species alternative synonymous codons occur at frequencies that are clearly nonrandom (1,2). The pattern of bias is broadly similar in genes from a single species or taxonomic group (3), but there is also considerable within-species heterogeneity. In the two best studied organisms, namely Escherichia coli (4,5) and the yeast Saccharomyces cerevisiae (6,7), genes differ largely in the degree of bias rather than its direction, and there is a strong positive correlation between the strength of bias and the level of gene expression. In both species, highly expressed genes have high frequencies of those codons thought to be optimally translated by the most abundant tRNA species (1,6,8-10), and this is most easily interpreted as the result of natural selection. Few other species have been investigated and, for several reasons, it is not clear to what extent these observations may be generalised. For example, prokaryotic genomes vary widely in overall G+C composition and codon usage patterns seem to reflect this (11), though whether this arises through selection or mutational biases is unclear. Mammalian genes also differ greatly in G+C composition at synonymous

sites in codons, apparently correlated with variations in local genomic G+C content (12,13), but also related (in some cases) to the tissue of gene expression (14).

Codon usage in Bacillus subtilis is, then, of interest for several reasons. Among prokaryotes only E.coli (4,5), and other genomes sharing a similar pattern of codon usage, i.e., some coliphages (15,16), and the closely related enteric bacterium Salmonella typhimurium (1), have been examined in any detail. B.subtilis is a gram positive bacterium unrelated to E.coli, and has a different G+C content, and so might be expected to exhibit a quite different pattern of codon usage. Indeed, preliminary investigations reveal this difference, but interestingly also suggest that codon usage in B.subtilis differs in several aspects which could not have been predicted. Firstly, it seems that there is comparatively little bias in B.subtilis genes (17). Secondly, one of the stronger elements of codon usage bias in E.coli and S.cerevisiae, namely an apparent preference for codons with an intermediate codon-anticodon interaction energy (10), may be absent in Bacillus (18). Finally, the preliminary data also showed no evidence of the excess of RNY (R = purine, N = any base, Y = pyrimidine) codons seen in most other organisms (19).

B.subtilis is important both as a system for the investigation of genetics and development (20), and as a potential host for heterologous gene expression (21). Codons which appear only rarely in E.coli genes are known to reduce the rate of translation in vivo (22,23), and insertion of these "nonoptimal" codons into a gene can reduce its level of expression (24,25). Thus it may be useful to identify any nonoptimal B.subtilis codons. Also, knowledge of the "standard" pattern of codon usage is a prerequisite both for several methods for locating coding regions within DNA sequences (26,27), and in the optimal design of synthetic oligonucleotide probes (28).

Here, codon usage data have been compiled for 56 B.subtilis genes. A small set of highly expressed genes, with a high codon usage bias, have been identified. From this set of genes optimal and nonoptimal codons can be recognized, and the relative merits of alternative synonymous codons can be assessed. For each gene the extent of codon usage bias, specifically in the direction of the pattern seen in highly expressed genes, has been estimated, and genes expressed at alternative life cycle stages have been contrasted. The patterns of codon usage seen in B.subtilis genes are viewed as a balance between mutation and selection, and have general implications regarding the evolutionary basis of synonymous codon usage.

MATERIALS AND METHODS

The data set

56 B.subtilis gene sequences were examined (incomplete reading frames of less than 60 codons were excluded), comprising 15272 codons in total. The source of each sequence is indicated in Table 1.

Correspondence analysis

Correspondence analysis was performed using the Cornell Ecology FORTRAN Program DECORANA, written by M.O. Hill. No transformations of the data were performed other than reciprocal averaging.

Relative synonymous codon usage (RSCU)

Relative synonymous codon usage values are estimated as the ratio of the observed codon usage to that value expected if there is uniform usage within synonymous groups (7)

Scaled chi-square

Chi-square values were calculated for each gene, using uniform synonymous codon usage as the expectation (and thus yielding 41 degrees of freedom). Since these values are generally highly correlated with gene length, they were then scaled by division by the number of codons in the gene (excluding Trp and Met codons, which do not contribute to the chi). This approach has previously been found to be useful for normalizing a G statistic (related to chi) of codon usage for yeast genes (7), so that comparisons among genes of different lengths are possible.

Codon adaptation index (CAI)

This index has previously been described for E.coli and yeast (29). Codon usage in a reference set of highly expressed genes is used to estimate the "relative adaptiveness", w, of each codon. w is calculated as the frequency of use of that codon (in the reference set) relative to the frequency of the optimal codon for that amino acid. The CAI for any gene is estimated as the geometric mean of the w values corresponding to each of the codons in that gene (excluding Met, Trp and termination codons.) The maximum possible value for CAI is 1.0, when only optimal codons are used.

P2 index

This index (4) describes the proportion of codons conforming to the intermediate strength of codon-anticodon interaction energy rule of Grosjean and Fiers (10): P2 = (WWC + SSU) / (WWY + SSY),

where W = A or U, S = G or C, Y = C or U, and (for example) WWC is the observed number of codons of that description. Thus, for a gene with uniform codon usage P2 = 0.5.

Table 1. Codon bias indices for 56 B.subtilis genes.

Gene	Group	Codons	CAI	P2	G+C	chi	(chiE)	Ref
sspA	HI	69	0.85	0.85	0.34	0.58		(32)
rpmA	HI	94	0.80	0.71	0.39	0.50		(33)
sspB	HI	67	0.78	0.75	0.36	0.51		(32)
rpsK	HI	131	0.74	0.76	0.30	0.66		JBa 168:65
sspC	HI	72	0.74	0.67	0.19	0.54		JBa 161:333
rpmH	HI	44	0.67	0.56	0.39	0.34	(1.04)	NAR 13:2251
sacB	EE	473	0.56	0.55	0.42	0.33		(45)
orf63 *	OT	63	0.54	0.30	0.47	0.22		(33)
sspD	HI	64	0.53	0.50	0.42	0.12		(32)
rpoD	OT	371	0.50	0.51	0.38	0.25	(0.72)	(38)
gyrA	OT	821	0.50	0.66	0.44	0.20		NAR 13:2251
bgl	EE	242	0.48	0.49	0.38	0.16		NAR 12:5355
0.3 kb	OT	61	0.47	0.44	0.43	0.15		JMB 176:333
gyrB	OT	638	0.47	0.50	0.46	0.18		NAR 13:2251
aprE	EE	381	0.47	0.54	0.39	0.31		JBa 158:411
nprE	EE	521	0.47	0.55	0.44	0.20		JBa 160:15
rpoA *	OT	64	0.46	0.73	0.44	0.40		JBa 168:65
orf71	OT	71	0.46	0.25	0.43	0.26		NAR 13:2251
spoVAC	SP	150	0.46	0.45	0.36	0.22		JGM 131:1091
ISP-I	OT	319	0.46	0.48	0.48	0.16		JBa 167:110
dnaN	OT	378	0.46	0.43	0.36	0.27	(0.54)	NAR 13:2251
spoIIAB	SP	146	0.45	0.39	0.42	0.22		JGM 130:2147
trpC	OT	250	0.44	0.51	0.38	0.20	(0.24)	Gen 34:169
trpB	OT	400	0.43	0.41	0.41	0.16	(0.42)	Gen 34:169
spoVAF*	SP	81	0.43	0.35	0.46	0.22		JGM 131:1091
dnaA	OT	446	0.43	0.39	0.38	0.20	(0.33)	NAR 13:2251
gdh	OT	260	0.42	0.47	0.48	0.19		JBa 166:238
citG	OT	462	0.42	0.56	0.51	0.17		NAR 13:131
hisH *	OT	108	0.41	0.58	0.50	0.13		Gen 34:169
spoIIAC	SP	195	0.41	0.52	0.50	0.14		JGM 130:2147
purF	OT	476	0.40	0.49	0.54	0.20		JBC 258:10586
trpA	OT	267	0.40	0.49	0.47	0.14	(0.34)	Gen 34:169
sdhA	OT	202	0.40	0.50	0.45	0.15	(0.69)	JBa 166:1067
spo0F	SP	124	0.40	0.43	0.60	0.29		NAR 14:1063
P23	OT	196	0.40	0.31	0.33	0.27		NAR 14:4293
trpD	OT	337	0.39	0.45	0.44	0.15	(0.46)	Gen 34:169
dnaE	OT	603	0.39	0.42	0.44	0.13	(0.17)	JBC 260:3368
trpE	OT	515	0.39	0.43	0.44	0.12	(0.39)	Gen 34:169
spoVAD	SP	338	0.39	0.46	0.43	0.16		JGM 131:1091
pyrB	OT	304	0.39	0.41	0.49	0.19		JBC 261:11156
trpF	OT	215	0.38	0.31	0.44	0.15		Gen 34:169
amyE	EE	660	0.37	0.41	0.42	0.14		NAR 11:237
recF	OT	323	0.37	0.45	0.50	0.15	(0.27)	NAR 13:2251
spoIIAA	SP	119	0.37	0.36	0.49	0.20		JGM 130:2147
spoIIG	SP	239	0.37	0.42	0.47	0.13		Nat 312:376
0.5kb	OT	173	0.36	0.27	0.47	0.15		PNA 80:658
spo0A	SP	239	0.36	0.37	0.51	0.24		PNA 82:2647
spo0B	SP	192	0.36	0.25	0.44	0.31		PNA 81:7012
spoVAB	SP	141	0.36	0.48	0.50	0.17		JGM 131:1091

Table 1 (cont.)

dalR	OT	389	0.35	0.38	0.50	0.14	B/T 3:1003
gerA	OT	480	0.35	0.38	0.51	0.16	Gen 38:95
orf85 *	OT	85	0.35	0.38	0.55	0.09	NAR 12:5355
spoVAE	SP	323	0.35	0.37	0.48	0.18	JGM 131:1091
spoIID	SP	343	0.34	0.33	0.46	0.18	JGM 132:341
spoVE	SP	293	0.34	0.43	0.55	0.21	JGM 132:1883
spoVAA	SP	200	0.33	0.38	0.46	0.21	JGM 131:1091

Groups of genes: HI = ribosomal protein, and SASP genes; EE = extracellular enzyme genes; OT = "other" genes; SP = spo genes. * incomplete sequences.
orf63, orf71 and orf85 are unidentified reading frames.
CAI = Codon Adaptation Index.
P2 = Use of intermediate energy codons (see text).
G+C = fraction of G+C at silent sites in codons.
chi = Chi-square for deviation from equal synonym use, scaled by gene length (chiE value is for an E.coli homolog, where available).
Ref: References for B.subtilis sequence data (E.coli data from GenBank). Abbreviations: B/T = Bio/Technology; Gen = Gene; JBa = J.Bacteriol.; JBC = J.Biol.Chem.; JGM = J.Gen.Micro.; NAR = Nucleic Acids Res.; Nat = Nature; PNA = Proc. Natl. Acad. Sci. USA.

Dinucleotide frequencies

Positional dinucleotide frequencies are expressed as the ratio of observed to expected, where expected frequencies are derived from observed positional nucleotide frequencies. For example, the expected frequency of ApC at position 2:3 is obtained from A2 (the frequency of A at codon position 2) and C3.

An intraclass correlation, ICC, (30) was used to estimate similarity among arrays of dinucleotide frequencies, after log transformation. The ICC has a value of 1.0 when there is perfect correlation.

RESULTS

Correspondence analysis reveals differences among B.subtilis genes

Correspondence analysis is a multivariate statistical technique which has been used extensively to examine differentiation among genes in codon usage (3,16,31). Genes can be displayed on a plot where the two axes depict the first and second most important factors of dispersion (i.e., trends through the data). The distances between genes on this plot are then a reflection of their dissimilarity in codon usage with respect to these trends. In E.coli and yeast, ribosomal protein genes are very highly expressed and are among those with the highest bias in codon usage (4,7). On a correspondence analysis plot of E.coli genes, those which are highly

expressed are spatially distinguished (see, e.g., ref. 31). Similarly, a plot of the 56 B.subtilis genes appears to differentiate between genes according to their expression levels (Figure 1). The three ribosomal protein genes in the B.subtilis data set are markedly displaced from the origin (the average codon usage for all genes) in Figure 1. The displacement is principally on the first (horizontal) axis, and similar values on that axis are seen for several ssp (small, acid-soluble spore protein, SASP) genes, which are very highly expressed during sporulation (32). Genes encoding extracellular enzymes, which are nonconstitutively highly expressed, are also generally displaced from the origin in the same direction (with respect to the first axis) as the ribosomal protein genes (Figure 1). Genes whose function are required for sporulation (spo genes), appear at the other extreme on the first axis, but not far from the origin. Many may be regulatory genes, and expression levels are low, where known (33). Correspondence analyses carried out with (i) relative synonymous codon usage (RSCU) values, or (ii) amino acid composition, as input (not shown) demonstrate that differentiation on the first axis is not due to either gene length or amino acid composition, but strongly suggests that the latter factor contributes to the second axis. Various measures of bias calculated for each of the genes, and compilations of codon usage for groups of genes (see below) suggest that the first axis of the correspondence analysis discriminates largely on degree of bias in synonymous codon usage.

Codon usage is less biased in B.subtilis than in corresponding E.coli genes

Chi-square values estimating deviation from uniform synonymous codon usage, and scaled by gene length, have been calculated for each B.subtilis gene, and also for 11 E.coli genes (Table 1) which appear to be homologous to genes in the B.subtilis data set. In Table 1 highly expressed genes are seen to be the most highly biased, although we note that this statistic may compound biases in codon usage due to both selection and mutation. Comparing genes across the two species, the scaled chi-square values are lower (in every case) for the B.subtilis genes than their E.coli homologues, demonstrating that bias in synonymous codon usage is indeed lower in B.subtilis than in E.coli.

Codon usage in four groups of B.subtilis genes

We had categorised genes according to their gene products (see Table 1). From the correspondence analysis (Figure 1) we conclude that genes within these groups are broadly similar with respect to the major features of their codon usage, and also that two of these groups (i.e., the ribosomal protein

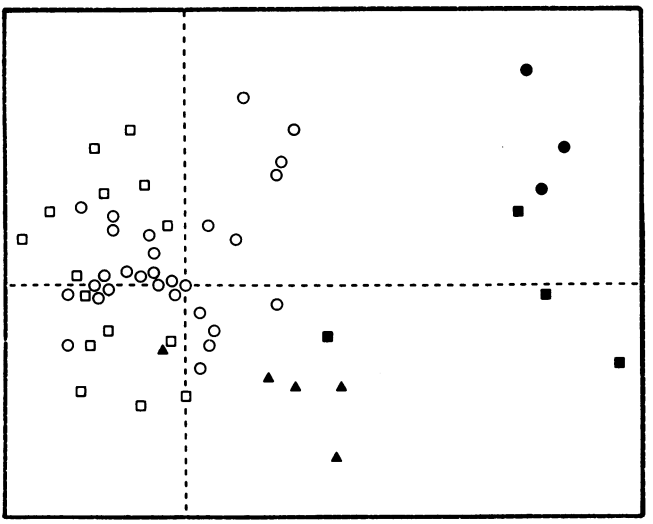

Figure 1. Correspondence analysis of codon usage in B.subtilis genes. The horizontal and vertical axes are the first and second axes, respectively. ● ribosomal protein genes, ■ SASP genes, ▲ extracellular enzyme genes, □ spo genes, ○ others.

genes and the SASP genes) are similar. Therefore, we have pooled codon usage values for four groups of genes (Table 2). In E.coli and yeast, RSCU values most deviant from 1.0 (a value which would indicate no bias) are seen in very highly expressed genes (5,7). Similarly, among the four groups of B.subtilis genes in Table 2, the 'very highly expressed' group has the most biased codon usage, although the total number of codons in that category is rather small. In E.coli, groups of genes can be identified which yield, for most sets of synonymous codons, a consistent trend in codon usage, from the very high bias in very highly expressed genes, to a much lower bias in genes (e.g., those encoding regulatory proteins) expressed at low levels (5). Again, in B.subtilis consistent trends in codon usage can be identified for most amino acids (Table 2). These trends seem to coincide with the first axis of the correspondence analysis, and among the four groups, that comprising spo genes (many of which may be regulatory and lowly expressed) appears to have the pattern of codon usage most dissimilar to the highly expressed genes. The spo genes do not, however, have less bias per se than many of the other genes, as can be seen most clearly from the group mean chi-square values (Table 3). Indeed, for some amino acids (e.g., Phe, Tyr) the spo genes have quite biased RSCU values (Table 2).

Table 2. Codon Usage in B.subtilis genes.

	HI BIAS	EXT ENZ	OTHERS	spo GENE	w
Phe TTT	8 0.73	50 1.16	227 1.32	104 1.55	0.571
TTC	14 1.27	36 0.84	118 0.68	30 0.45	1.000
Leu TTA	14 2.33	42 1.64	183 1.27	50 0.93	1.000
TTG	0 0.00	24 0.94	114 0.79	45 0.84	0.036
Leu CTT	12 2.00	36 1.40	220 1.52	70 1.30	0.857
CTC	2 0.33	11 0.43	94 0.65	41 0.76	0.143
CTA	7 1.17	8 0.31	50 0.35	20 0.37	0.500
CTG	1 0.17	33 1.29	207 1.43	97 1.80	0.071
Ile ATT	7 0.95	52 1.46	327 1.48	121 1.38	0.500
ATC	14 1.91	48 1.35	265 1.20	111 1.27	1.000
ATA	1 0.14	7 0.20	73 0.33	31 0.35	0.071
Met ATG	9 1.00	37 1.00	216 1.00	91 1.00	---
Val GTT	16 1.68	44 1.27	182 1.13	59 1.03	1.000
GTC	3 0.32	35 1.01	171 1.07	78 1.36	0.188
GTA	12 1.26	32 0.92	144 0.90	44 0.77	0.750
GTG	7 0.74	28 0.81	145 0.90	48 0.84	0.438
Ser TCT	24 3.25	58 1.63	128 1.31	29 0.88	1.000
TCC	0 0.00	28 0.78	78 0.80	17 0.52	0.021
TCA	11 1.53	37 1.04	124 1.27	47 1.42	0.458
TCG	0 0.00	19 0.53	58 0.59	29 0.88	0.021
Pro CCT	7 2.15	36 1.80	102 1.22	26 0.99	1.000
CCC	0 0.00	2 0.10	31 0.37	8 0.30	0.071
CCA	5 1.54	23 1.15	57 0.68	29 1.10	0.714
CCG	1 0.31	19 0.95	144 1.72	42 1.60	0.143
Thr ACT	15 1.94	32 0.75	80 0.61	21 0.49	1.000
ACC	0 0.00	16 0.37	83 0.64	37 0.87	0.033
ACA	13 1.68	84 1.96	212 1.63	69 1.61	0.867
ACG	3 0.39	39 0.91	146 1.12	44 1.03	0.200
Ala GCT	40 2.81	50 1.00	176 1.02	54 0.93	1.000
GCC	1 0.07	41 0.82	129 0.75	62 1.06	0.025
GCA	11 0.77	67 1.35	201 1.16	63 1.08	0.275
GCG	5 0.35	41 0.82	185 1.07	54 0.93	0.125
Tyr TAT	1 0.67	67 1.18	181 1.23	63 1.34	0.500
TAC	2 1.33	47 0.82	113 0.77	31 0.66	1.000
TER TAA	6 -	4 -	18 -	9 -	---
TAG	1 -	0 -	1 -	2 -	---
His CAT	6 2.00	30 1.25	132 1.25	42 1.29	1.000
CAC	0 0.00	18 0.75	80 0.75	23 0.71	0.083
Gln CAA	28 1.65	67 1.38	183 1.03	48 0.89	1.000
CAG	6 0.35	30 0.62	174 0.97	60 1.11	0.214

	HI BIAS	EXT ENZ	OTHERS	spo GENE	w
Asn AAT	10 0.59	72 0.89	204 1.07	66 1.20	0.417
AAC	24 1.41	90 1.11	179 0.93	44 0.80	1.000
Lys AAA	31 1.82	124 1.59	461 1.47	144 1.36	1.000
AAG	3 0.18	32 0.41	166 0.53	68 0.64	0.097
Asp GAT	5 0.59	84 1.22	312 1.24	105 1.25	0.417
GAC	12 1.41	54 0.78	191 0.76	63 0.75	1.000
Glu GAA	17 1.42	65 1.49	536 1.36	132 1.41	1.000
GAG	7 0.58	22 0.51	254 0.64	55 0.59	0.412
Cys TGT	0 1.00	1 0.67	30 0.94	18 1.29	1.000
TGC	0 1.00	2 1.33	34 1.06	10 0.71	1.000
TER TGA	0 -	1 -	7 -	3 -	---
Trp TGG	1 1.00	35 1.00	47 1.00	22 1.00	---
Arg CGT	23 2.88	16 1.66	105 1.39	18 0.85	1.000
CGC	14 1.75	11 1.14	85 1.13	21 0.99	0.609
CGA	0 0.00	8 0.83	41 0.54	13 0.61	0.022
CGG	1 0.13	3 0.31	59 0.78	21 1.13	0.043
Ser AGT	3 0.42	18 0.50	59 0.60	21 0.64	0.125
AGC	5 0.70	54 1.51	140 1.43	55 1.67	0.208
Arg AGA	10 1.25	12 1.24	125 1.66	38 1.80	0.435
AGG	0 0.00	8 0.83	37 0.49	13 0.61	0.022
Gly GGT	21 1.38	46 0.98	157 0.97	32 0.53	0.955
GGC	17 1.11	72 1.54	213 1.31	73 1.22	0.773
GGA	22 1.44	56 1.20	193 1.19	83 1.38	1.000
GGG	1 0.07	13 0.28	87 0.54	52 0.87	0.045

Codon usage is presented as the number of codons, and the Relative Synonymous Codon Usage.
Groups of genes defined in Table 1; total number of codons: HI BIAS (541), EXT ENZ (2307), OTHERS (3123), spo GENE (9301).
w "relative adaptedness" of each codon, derived from HI BIAS group, and used in calculation of CAI (see MATERIALS AND METHODS). Note that the usage of any codon absent in the HI BIAS group is arbitrarily adjusted to 0.5, to avoid zero values for w and the CAI (29).

The frequency of alternative synonymous codons in a reference set of very highly expressed genes can be used to quantify the relative fitness of each of those codons, and the Codon Adaptation Index (CAI) of any gene is then the geometric mean of the "fitness" values corresponding to each of the codons in that gene (29). Here we have used the seven very highly expressed genes as a B.subtilis reference set, and CAI values for the 56 genes are shown in Table 1. These CAI values are highly correlated with the first axis

Table 3. Mean codon bias indices (see Table 1 for key)

Group	genes	CAI	P2	G+C	chi
HI BIAS	7	0.73	0.69	0.34	0.46
EXT ENZ	5	0.47	0.51	0.41	0.23
OTHERS	29	0.42	0.45	0.45	0.18
spo GENE	15	0.38	0.40	0.48	0.21

of the correspondence analysis. As expected, the five genes identified from Figure 1 as having the most biased codon usage also have the highest CAI values. Genes encoding extracellular enzymes have generally high values, with the exception of amyE. Sporulation genes have generally low CAI values. Consideration of the mean CAI values for the four groups of genes (Table 3) shows a clear trend across groups. From the CAI values it can be seen that spo genes have the least amount of bias towards those codons favoured in the highly expressed genes.

(CAI values for the B.subtilis genes are similar to, or often higher than, CAI values for E.coli genes [calculated using a reference set of very highly expressed E.coli genes], despite the generally lower bias seen in B.subtilis. This apparent anomaly arises because the reference set of highly expressed B.subtilis genes has comparatively low codon bias.)

Choice among pairs of codons translated by the same tRNA

Grosjean and Fiers (10) have suggested that among codons which are either A/T- or G/C-rich at positions one and two, and which end in a pyrimidine, there is selection for that base (T or C) which yields the less extreme overall codon-anticodon interaction energy. Gouy and Gautier (4) have devised an index (P2) of this preference, calculated as the frequency of those codons predicted to be preferred. In E.coli and yeast, highly expressed genes have high P2 values (0.7-0.9) indicating a strong preference, while other genes have values close to 0.5 indicating little preference (4,7). P2 values for the 56 B.subtilis genes are shown in Table 1. A few genes have P2 values greater than 0.5. These are predominantly those in the highly expressed group, and consideration of mean values for the groups of genes reveals a trend echoing that for the CAI (Table 3). However, in a comparison of highly expressed genes, P2 values from B.subtilis are lower than those from E.coli or S.cerevisiae.

Base composition and dinucleotide frequencies

If nucleotide sites can be inferred to be under very weak selective

Table 4. Base composition in B.subtilis genes.

	Pos	HI BIAS	EXT ENZ	OTHERS	spo GENE
T	3	0.37	0.30	0.28	0.27
C	3	0.20	0.25	0.22	0.23
A	3	0.34	0.28	0.28	0.26
G	3	0.08	0.17	0.22	0.24
R	1	0.65	0.65	0.65	0.65
R	3	0.43	0.45	0.50	0.50
Y	1	0.35	0.35	0.35	0.35
Y	3	0.57	0.55	0.50	0.50

Groups of genes as in Table 1. R = A or G, Y = C or T.
Pos = base position within the codon.

constraint, then base composition and dinucleotide frequencies may reflect the direction of mutational pressures. The frequency of G+C at synonymously variable positions in each gene has been calculated (Table 1). Overall genomic G+C content in B.subtilis is approximately 42% (34). There is a strong negative correlation between %G+C and CAI for the 56 genes, and this is echoed in a trend in G+C content across the four groups of genes (Table 3). It appears that selection for particular synonymous codons, most evident in the highly expressed genes, has the effect of reducing the G+C content at synonymously variable codon positions. Base composition at positions 1 and 2 does not vary much between the four groups of genes (data not shown). Overall, about 65% of codons begin with a purine. From the base composition at position 3 (Table 4) it can be seen that the low G+C content at synonymously variable sites in the highly expressed genes is largely due to a scarcity of G-ending codons. Across the four groups of genes G3 (G at position 3) increases as A3 and, to a lesser extent, T3 decrease. The low frequency of G3 in more highly expressed genes leads to a small excess of pyrimidine-ending codons in those genes. In the 15 sporulation and 29 "other" genes purines and pyrimidines are at equal frequencies in position 3.

There are three dinucleotide positions with respect to a coding sequence, i.e., 1:2, 2:3 and 3:1, denoting the codon positions of the constituent bases. Base composition at positions 1 and 2, and the 1:2 dinucleotide frequencies, are largely determined by the amino acid composition of encoded proteins. If selection at the translational level is strong enough to overcome any selection at the nucleic acid structural level, as well as mutation and random drift, then the 2:3 and 3:1 dinucleotides reflect synonymous codon

Table 5. Positional dinucleotide frequencies.

	HI BIAS		EXT ENZ		OTHERS		spo GENE	
	2:3	3:1	2:3	3:1	2:3	3:1	2:3	3:1
TT	0.96	1.29	1.19	1.00	1.30	1.23	1.32	1.12
TC	1.34	1.11	1.04	1.05	1.15	1.02	1.17	1.00
TA	0.83	0.86	0.64	0.67	0.62	0.68	0.56	0.74
TG	1.02	0.92	1.25	1.28	0.94	1.16	0.95	1.18
CT	1.67	0.47	0.95	0.96	0.86	0.90	0.73	0.97
CC	0.04	0.94	0.57	0.83	0.74	0.95	0.83	0.92
CA	0.85	1.05	1.24	1.17	1.07	1.12	1.22	1.02
CG	0.99	1.23	1.40	0.94	1.39	0.97	1.27	1.04
AT	0.38	1.05	1.00	0.99	0.90	0.96	1.03	1.09
AC	1.20	0.90	1.02	0.76	0.80	0.69	0.73	0.61
AA	1.44	1.05	1.11	1.19	1.30	1.25	1.27	1.29
AG	1.58	1.00	0.74	0.94	0.94	0.97	0.92	0.92
GT	1.05	0.73	0.81	1.08	0.88	0.86	0.67	0.76
GC	1.51	1.07	1.69	1.68	1.55	1.48	1.44	1.58
GA	0.78	1.37	0.83	1.04	0.92	0.97	1.05	0.96
GG	0.25	0.80	0.53	0.60	0.68	0.83	0.90	0.83

Dinucleotide frequencies calculated as observed/expected, for the 59 synonymously variable codons only. 2:3 and 3:1 refer to dinucleotide position with respect to the reading frame.

preference and any codon "context" effects (35-37), respectively. Base composition at position 3 results from the net effect of these forces.

Dinucleotide frequencies, expressed relative to the expected occurrence of each dinucleotide (predicted from the frequencies of each base at each codon position), have been calculated for the four groups of genes (Table 5). An intraclass correlation (30) has been used to quantify similarities between arrays of dinucleotide frequencies. The two large groups of more lowly expressed genes are very similar with respect to dinucleotide composition: the 2:3 and 3:1 frequencies across the two groups are highly correlated (ICC values around 0.9). The 2:3 and 3:1 dinucleotide frequencies within each group are also highly correlated (ICC values around 0.8), for all but the highly expressed group of genes (where there is a nonsignificant negative correlation). While the frequencies from the latter group are certainly subject to rather large random errors because of the small sample size, nevertheless the difference between this group and the others is so large that it is likely to be real. The dinucleotide frequencies in the highly expressed group are, of course, influenced by the (inferred) selection of

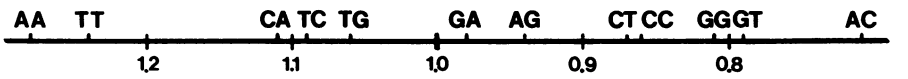

Figure 2. Frequencies of complementary dinucleotides. (Frequencies have been averaged over the 2:3 and 3:1 positions of the SP and OT groups [see Table 1] of genes; frequencies are expressed as observed/expected.)

synonymous codons in those genes. The similarity of the 2:3 and 3:1 dinucleotide frequencies in the other groups of genes suggests a relative lack of translational selection. This inference is strongly endorsed by a very high correlation (ICC = 0.96) between the frequencies of complementary dinucleotides, e.g., TT and AA; AC and GT (Figure 2). Such a correlation suggests that the observed preferences are a characteristic of the double-stranded DNA rather than the single-stranded message, and might arise from selection for DNA structure, or more simply from mutational biases in the absence of selection.

DISCUSSION

The biological basis of the pattern of synonymous codon usage and, in particular, its variability among genes within the same species, has been satisfactorily examined for two species only, namely E.coli and S.cerevisiae. It has been suggested that the pattern of codon usage in B.subtilis differs in several fundamental ways (17,18), and indeed the present analysis of 56 B.subtilis genes confirms some of these preliminary observations. For example, there is less overall bias in codon usage in B.subtilis, and those codons that are "preferred" are generally different from those in the other species. However, in several significant ways codon usage in B.subtilis can be seen to fit into the same pattern as observed in E.coli and S.cerevisiae. Thus these data provide evidence suggesting the generality of certain phenomena, at least among unicellular organisms.

Strength of codon preference is correlated with gene expression level

We have distinguished four groups of genes, on biological criteria, and found that these groups differ in their degree of codon usage bias, as illustrated by the mean values of the various codon usage indices for each of the groups (Table 3). A highly biased, and atypical, pattern of codon usage is found in one small group, comprising ribosomal protein genes, which are expected to be very highly expressed, and SASP genes which are known to be so. CAI values for B.subtilis genes, measuring the extent of bias towards this preferred pattern of synonymous codon usage, are highly correlated with

the known or expected expression levels of those genes. For example, the products of extracellular enzyme genes can be quite abundant, and these genes have comparatively high CAI values. Also, rpoD has a higher than average CAI value, and its gene product is present at 2,000-10,000 protein molecules per cell (38). In contrast, the products of spo0A and spo0B are known to be present at very low levels (10-100 protein molecules per cell, Ref.33) and these genes have low CAI values (Table 1).

It is most likely that the codon biases observed in highly expressed B.subtilis genes have resulted from selection at the level of translation. Certain codons may be more efficiently translated if they correspond to more abundant tRNAs and/or because they bind to the anticodon more efficiently than other codons recognised by the same tRNA. Unfortunately, the relative abundances of different tRNA species in B.subtilis are unknown, but one indicator is the number of genes that encode the tRNA. For most of the cases where a B.subtilis tRNA species is known to be encoded by more than one gene (39) the codon translated (without wobble) is a preferred codon in the highly expressed group of genes (Table 2). Among pairs of codons translated by the same tRNA, e.g., for amino acids encoded by only two triplets, the (unmodified) tRNA has either G or U in the first position of the anticodon, and again those codons whose translation does not require wobble binding are preferred in the highly expressed group (Table 2). Thus, while there is some evidence in highly expressed B.subtilis genes of the third base pyrimidine bias effect of Grosjean and Fiers (10), this may result largely from selection against wobble (40).

We note that it seems more likely that, generally, codon usage is modulated by long-term gene expression level, than that codon usage can modulate gene expression (5,31). Holm (31) has suggested that natural selection probably differentiates between synonymous codons according to the cost of proof-reading during translation. The more abundant the gene product, over the entire life cycle, the greater the selective difference between alternative codons in the gene. Then genes which are highly expressed, but only rarely, would not exhibit very high codon bias.

In the highly expressed B.subtilis genes certain codons are strongly avoided and, by analogy with E.coli (24,25), it is possible that the presence of such codons could reduce the yield of a heterologous gene product in B.subtilis, particularly if those nonoptimal codons are clustered (41). However, there are now a large number of instances where high levels of

heterologous gene expression have been achieved despite the presence of many nonoptimal codons (e.g., 42,43).

Why do B.subtilis genes have low codon bias?

The relatively low levels of synonymous codon usage bias in B.subtilis (compared to E.coli or S.cerevisiae) could arise either because selective differences between codons are smaller, or because selection is less effective. In both E.coli and S.cerevisiae tRNA abundance seems to be a major selective factor (1). Ogasawara (17) has suggested that since the tRNA genes in B.subtilis are found in major gene clusters, and are transcriptionally co-regulated (39), there may be comparatively small differences in relative abundance among different tRNAs (and hence smaller selective differences between codons.) However, since preferences among codons recognised by the same tRNA (as estimated, for example, by the P2 statistic) are also less pronounced in B.subtilis than in E.coli, it is likely that the lower bias is not a consequence of tRNA abundances alone. It is possible that natural selection is generally less efficacious in discriminating between synonymous codons in B.subtilis. This might occur if the effective population size is smaller, or if the mutation rate is higher, than in other species (44).

Codon preferences and differentiation

B.subtilis is a sporulating organism. It differentiates temporally through a cycle of germination, vegetative growth, and sporulation. It is interesting to consider whether codon preferences differ at these various stages, possibly in relation to tRNA modifications or changes in tRNA abundance (39). For example, it has been suggested that the different patterns of codon usage seen in sacB and amyE may have arisen through selection, because the former is expressed during vegetative growth, while the latter is expressed during sporulation (45). We have not detected any such effects. Codon usage in the ribosomal protein genes, which are probably expressed throughout the life-cycle, closely resembles that of the SASP genes, which are expressed in the forespore during sporulation only. No obvious differences in codon usage can be seen among genes essential to different stages of sporulation (e.g., spoII and spoV). The difference between amyE and sacB seems to be in degree of codon bias, with selection apparently more effective in sacB. This may reflect changes in the strength of selection during the life cycle.

Codon usage in lowly expressed genes reveals mutational biases

It is likely that synonymous codon usage in genes expressed at low levels is subject to little, if any, translational selection (5). This is

confirmed by the pattern of codon usage in the group of sporulation genes. In those genes there is an apparent "preference" for a different set of codons compared to the very highly expressed genes (Table 2). However, consideration of dinucleotide frequencies, and particularly the correlation of the frequencies of complementary dinucleotides, strongly suggests that these preferences do not result from translational selection. Rather, in those genes codon usage is more likely to be influenced by mutation pressures, or by selection operating at the level of DNA structure. Note that biases in mutation patterns can be responsible for such skewed dinucleotide frequencies since neighbouring bases can affect mutation rates (e.g., 46).

Among single-celled organisms there is a general correlation between genomic G+C content and the G+C composition at synonymous sites in genes (11, 47). The genomes of B.subtilis and S.cerevisiae are similar in being A/T rich, with G+C content in the region of 40-42% (34). Interestingly, patterns of codon usage in these two organisms are rather different, particularly in highly expressed genes. In S.cerevisiae stronger codon preference increases the frequency of G+C at synonymous sites (7), whereas in B.subtilis highly expressed genes have an even lower G+C content than the rest of the genome (Table 5). Mutational bias (inferred from the lowly expressed genes) tends to raise G+C above the level reported for the whole B.subtilis genome. However, this bias does not lead to a general excess of Y-ending codons. The base composition of the highly expressed genes suggests that the preponderance of RNY codons is perhaps best viewed (here, and in other organisms) as a composite of two separate phenomena, i.e., an excess of codons beginning in R (which may be a direct consequence of amino acid usage), and an excess of codons ending in Y (an artefact of the particular codon preference in different organisms.)

Codon usage can be a useful criterion for establishing whether open reading frames are truly genes. The absence of an excess of Y-ending codons suggests that the "RNY method" (48) will not be particularly useful for finding genes in B.subtilis DNA sequences. Furthermore, the comparative lack of bias in a large proportion of the genes examined here suggests that methods dependent on a bias in codon usage in the gene being located (26,27) are also likely to be less successful than in other species.

ACKNOWLEDGEMENTS

We are grateful to D.G. Higgins for assistance with correspondence analysis, and to B.A. Cantwell, K.M. Devine and F.G. Wright for reading a draft of the manuscript. This work was carried out using the facilities of the Irish National Centre for BioInformatics.

*To whom correspondence should be addressed

REFERENCES

1. Ikemura, T. (1985) Mol. Biol. Evol. 2, 13-34.
2. Maruyama, T., Gojobori, T., Aota, S. and Ikemura, T. (1986) Nucleic Acids Res. 14, r151-197.
3. Grantham, R., Gautier, C., Gouy, M., Mercier, R. and Pave, A. (1980) Nucleic Acids Res. 8, r49-r62.
4. Gouy, M. and Gautier, C. (1982) Nucleic Acids Res. 10, 7055-7074.
5. Sharp, P.M. and Li, W-H. (1986) Nucleic Acids Res. 14, 7737-7749.
6. Bennetzen, J.L. and Hall, B.D. (1982) J. Biol. Chem. 257, 3026-3031.
7. Sharp, P.M., Tuohy, T.M.F. and Mosurski, K.R. (1986) Nucleic Acids Res. 14, 5125-5143.
8. Ikemura, T. (1981) J. Mol. Biol. 146, 1-21.
9. Ikemura, T. (1982) J. Mol. Biol. 158, 573-597.
10. Grosjean, H. and Fiers, W. (1982) Gene 18,199-209.
11. Bibb, M.J., Findlay, P.R. and Johnson, M.W. (1984) Gene 30, 157-166.
12. Bernardi, G., Olofsson, B., Filipski, J., Zerial, M., and Salinas, J., Cuny, G., Meunier-Rotival, M. and Rodier, F. (1985) Science 228, 953-958.
13. Aota, S-i. and Ikemura, T. (1986) Nucleic Acids Res. 14, 6345-6355.
14. Newgard, C.B., Nakano, K., Hwang, P.K. and Fletterick, R.J. (1986) Proc. Natl. Acad. Sci. USA 83, 8132-8136.
15. Sharp, P.M., Rogers, M.S. and McConnell, D.J. (1985) J. Mol. Evol. 21, 150-160.
16. Grantham, R., Greenland, T., Louail, S., Mouchiroud, D., Prato, J.L., Gautier, C. and Gouy, M. (1985) Bull. Inst. Pasteur 83, 95-148.
17. Ogasawara, N. (1985) Gene 40, 145-150.
18. McConnell, D.J., Cantwell, B.A., Devine, K.M., Forage, A.J., Laoide, B.M., O'Kane, C., Ollington, J.F. and Sharp, P.M. (1986) Ann. N.Y. Acad. Sci. 469, 1-17.
19. Shepherd, J.C.W. (1984) Trends Biochem. Sci. 9, 8-10.
20. Dubnau, D.A., Ed. (1982) The Molecular Biology of the Bacilli, Vol 1, Academic Press, New York.
21. Doi, R.H., Wong, S-L. and Kawamura, F. (1986) Trends Biotech. 4, 232-235.
22. Pedersen, S. (1984) EMBO J. 3, 2895-2898.
23. Varenne, S., Buc, J., Lloubes, R. and Lazdunski, C. (1984) J. Mol. Biol. 180, 549-576.
24. Robinson, M., Lilley, R., Little, S., Emtage, J.S., Yarranton, G., Stephens, P., Millican, A., Eaton, M. and Humphreys, G. (1984) Nucleic Acids Res. 12, 6663-6671.
25. Bonekamp, F., Andersen, H.D., Christensen, T. and Jensen, K.F. (1985) Nucleic Acids Res. 13, 4113-4123.
26. Staden, R. (1984) Nucleic Acids Res. 12, 551-567.
27. Gribskov, M., Devereux, J. and Burgess, R.R. (1984) Nucleic Acids Res. 12, 539-549.
28. Lathe, R. (1985) J. Mol. Biol. 183, 1-12.
29. Sharp, P.M. and Li, W-H. (1987) Nucleic Acids Res. 15, 1281-1295.
30. Snedecor, G.W. and Cochran, W.G. (1967) Statistical Methods, 6th edn. pp. 294-295, Iowa State University Press, Ames, Iowa.
31. Holm, L. (1986) Nucleic Acids Res. 14, 3075-3087.
32. Connors, M.J., Mason, J.M. and P. Setlow (1986) J. Bacteriol. 166, 417-425.
33. Ferrari, F.A., Trach, K. and Hoch, J.A. (1985) J. Bacteriol. 161, 556-562.

34. Normore, W.M. (1973) in Handbook of Microbiology, Laskin, A.I. and Lechevalier, H.A. Eds., Vol.2, pp. 585-740, CRC, Cleveland, Ohio.
35. Yarus, M. and Folley, L.S. (1985) J. Mol. Biol. 182, 529-540.
36. Schpaer, E.G. (1986) J. Mol. Biol. 188, 555-564.
37. Gouy, M. (1987) Mol. Biol. Evol. 4, 426-444.
38. Gitt, M.A., Wang, L-F. and Doi, R.H. (1985) J. Biol. Chem. 260, 7178-7185.
39. Vold, B.S. (1985) Microbiol. Rev. 49, 71-80.
40. Fitch, W.M. (1976) Science 194, 1173-1174.
41. Varenne, S. and Lazdunski, C. (1986) J. Theor. Biol. 120, 99-110.
42. Tokunaga, T., Iwai, S., Gomi, H., Kodama, K., Ohtsuka, E., Ikehara, M., Chisaka, O. and Matsubara, K. (1985) Gene 39, 117-120.
43. Kniskern, P.J., Hagopian, A., Montgomery, D.L., Burke, P., Dunn, N.R., Hofman, K.J., Miller, W.J. and Ellis, R.W. (1986) Gene 46, 135-141.
44. Sharp, P.M. and Li, W-H. (1986) J. Mol. Evol. 24, 28-38.
45. Steinmetz, M., LeCoq, D., Aymerich, S., Gonzy-Treboul, G. and Gay, P. (1985) Mol. Gen. Genet. 200, 220-228.
46. Bulmer, M. (1986) Mol. Biol. Evol. 3, 322-329.
47. Bernardi, G. and Bernardi, G. (1985) J. Mol. Evol. 22, 363-365.
48. Shepherd, J.C.W. (1981) Proc. Natl. Acad. Sci. U.S.A. 78, 1596-1600.